Enhancing the Vision for Managing California's Environmental Information

Produced by the Environmental Data Summit Organizing Committee under the leadership of the Delta Stewardship Council's Delta Science Program

February 2015
Sacramento, California

Acknowledgements

The production of this vision document was a truly collaborative effort. The organizing team assembled by the Delta Stewardship Council's Delta Science Program included representatives from state and federal agencies, academia, nonprofits, and private companies. We owe a debt of gratitude to all who contributed their time, intelligence, creativity, fortitude, and open-mindedness to the effort.

This paper represents some of the best ideas to emerge from the Data Summit, convened in June 2014. The Summit organizing committee played an instrumental role, not only in conceiving and planning the conference, but also in ushering this vision document through its many phases.

The reviewers graciously lent their time to peruse the document and offer their editorial suggestions. Each offered an astute expert's eye to detail. Any remaining faults in the document are not attributable to gaps in their vision.

Many names will be omitted, but among those whose labors are keenly appreciated are the following:

Data Summit Organizers

Delta Stewardship Council ■ Peter Goodwin ■ Rainer Hoenicke ■ George Isaac ■ Garrett Liles	Department of Water Resources ■ David Harris ■ Rich Juricich ■ Nancy Miller
State and Federal Contractors Water Agency ■ Val Connor	Delta Conservancy ■ Shakoora Azimi-Gaylon
San Francisco Estuary Institute ■ Cristina Grosso ■ Tony Hale	State Water Resources Control Board ■ Karen Larsen ■ Jon Marshack
34 North ■ Dave Osti	Department of Fish and Wildlife ■ Gregg Erickson

Writing Team

Shakoora Azimi-Gaylon Assistant Executive Officer, Delta Conservancy
Stephanie Fong Senior Staff Scientist, State and Federal Contractors Water Agency
Peter Goodwin Lead Scientist, Delta Stewardship Council's Delta Science Program

Tony Hale, PhD Program Director, Environmental Informatics, San Francisco Estuary Institute
George Isaac Senior Environmental Scientist, Delta Stewardship Council's Delta Science Program
Amye Osti Founder/CEO, 34 North
Fraser Shilling, PhD Research Scientist, UC Davis Department of Environmental Science and Policy
Tad Slawecki Senior Engineer, LimnoTech, Ann Arbor, Michigan
Steve Steinberg, PhD Director, Information Management and Analysis, Southern California Coastal Water Research Project
Mark Tompkins, PhD Senior Engineering Geomorphologist, New Fields

Reviewers of this paper

Vladan Babovic Associate Professor, Department of Civil and Environmental Engineering National University of Singapore
Adam Ballard Staff Environmental Scientist, Department of Fish and Wildlife
Jon Bishop Chief Deputy Director, State Water Resources Control Board
Scott Cantrell Branch Chief, Department of Fish and Wildlife
Gary Darling Business Service Assistant, Department of Water Resources
J. Carl Dealy Project Manager, US Bureau of Reclamation
Greg Gearheart Deputy Director for the Office of Information Management and Analysis State Water Resources Control Board
Abdul Khan Supervising Engineer, Department of Water Resources
Lindsay Irving Visualization Project Manager
Tom Lupo Deputy Director, Data and Technology Division, Department of Fish and Wildlife
Ray McDowell FESSRO, Department of Water Resources
Nancy Miller Deputy Chief Information Officer, Department of Water Resources
Doug Parker

Director, California Institute for Water Resources
John Ryan Program Manager, Delta Stewardship Council
Dragan Savic Professor of Hydroinformatics, University of Exeter, United Kingdom
Stephani Spaar Chief, Office of Water Quality, DWR Division of Environmental Services
Alicia Torregrosa Physical Scientist, US Geological Survey

Table of Contents

Contents

Executive Summary

The Environmental Data Summit, convened under the auspices of the Delta Stewardship Council's Delta Science Program in June 2014, witnessed remarkable participation from experts across California, the nation, and even the world. Summit attendees from the public, private, federal, and non-profit sectors shared their views regarding the urgent needs and proposed solutions for California's data-sharing and data-integration challenges, especially pertaining to the subject of environmental resource management in the era of "big data." This is a time when our data sources are growing in number, size, and complexity. Yet our ability to manage and analyze such data in service of effective decision-making lags far behind our demonstrated needs.

> Our work is to establish a vision that can be broadly shared among agency, NGO, tribal, academic, and public stakeholders. By fostering principled consensus, we wish to demonstrate that California has a broad plan for addressing its environmental data management challenges, thereby fostering a collaborative, fundable package of solutions that can be implemented over a sustained period of time with minimal disruption to established processes.

In its review of the sustainability of water and environmental management in the California Bay-Delta, the National Research Council (NRC) found that "only a synthetic, integrated, analytical approach to understanding the effects of suites of environmental factors (stressors) on the ecosystem and its components is likely to provide important insights that can lead to enhancement of the Delta and its species" (National Research Council 2012). The present "silos of data" have resulted in "silos of science" and impeded our ability to make informed decisions. While resolving data integration challenges will not, by itself, produce better science or better natural resource outcomes, progress in this area will provide a strong foundation for decision-making. Various mandates ranging from the California Water Action Plan to the President's executive order demanding federal open data policies demonstrate the consensus on the merits of modern data sharing at the scale and function needed to meet today's challenges.

This white paper emerges from the Summit as an instrument to help identify such opportunities to enhance California's cross-jurisdictional data management. As a resource to policy makers, agency leadership, data managers, and others, this paper articulates some key challenges as well as proven solutions that, with careful and thoughtful coordination, can be implemented to overcome those obstacles. Primarily featured are tools that complement the State's current investments in technology, recognizing that success depends upon broad and motivated participation from all levels of the public agency domain.

This document describes examples, practices, and recommendations that focus on California's Delta as an opportune example likely to yield meaningful initial results in the face of pressing challenges. Once proven in the Delta, however, this paper's recommended innovations would conceivably be applied statewide in subsequent phases.

For the purposes of this executive summary, here we highlight some of the findings and recommendations found within the white paper. This subset should provide insight into some of the white paper's primary assertions. The full list of findings and recommendations follows this executive summary.

Findings

- The State's data-governance policies are lacking in definition and current application. A new governance framework -- a system of decision rights and accountability for information-related processes -- should be established that facilitates broader decisions and standards regarding the State's data management.
- "Transparency" is a fundamental attribute of public data, but its definition has changed with advances in technology. The public stakeholders and peer agencies alike now seek data on demand.
- Clear and careful documentation of data quality and data formats through metadata avoids misunderstandings and misapplication of information – increasing the effectiveness of management decisions, reducing disputes, and obviating some basis of litigation. Clear standards also help to promote compatibility among datasets for purposes of aggregation and analysis.
- Coordinated and collaborative data management must be conducted using business models that foster sustained, incremental investment and partnership with non-governmental partners.

Recommendations

- Data governance oversight: applying data standards, documenting data, and seeking strategic alliance with national and global initiatives
- Develop a data federation strategy with a specific, time-bound roadmap. This effort must complement the work of the data standards implementation.

Develop a business case and adopt a funding strategy in service of a sustainable business model.

Full Findings & Recommendations

The following is a collection of the findings and recommendations located in the white paper's three sections. For additional information and context on each item, please consult the full white paper.

Findings

In many key ways, California's technology infrastructure and approaches to problem-solving are "behind the curve" of the mounting challenges it faces with respect to natural resource management. Specific findings related to this assessment follow:

Evolving Data Stewardship

1. Data sharing is one of the most fundamental building blocks in effective scientific and resource-management collaboration.
2. The State's data-governance policies are lacking in definition and current application. A new governance framework -- a system of decision rights and accountability for information-related processes -- should be established that facilitates broader decisions and standards regarding the State's data management.
3. Innovative initiatives are already underway that make data accessible, understandable and shareable – and these efforts are already reaping significant rewards in terms of saved time, enhanced collaboration among different organizations, and accelerated knowledge discovery that provides better information to make decisions on California's ecology and water-supply challenges.
4. "Transparency" is a fundamental attribute of public data, but its definition has changed with advances in technology. The public stakeholders and peer agencies alike now seek data on demand.
5. Data used in decision-making are often aggregated or transformed. The "reproducibility" of any data transformations is a measure directly related to transparency. Expectations and needs outpace the current capabilities to deliver such data stewardship information to the public and interested agencies.
6. Clear and careful documentation of data quality and data formats through metadata avoids misunderstandings and misapplication of information – increasing the effectiveness of management decisions, reducing disputes, and obviating some basis of litigation. Clear standards also help to promote compatibility among datasets for purposes of aggregation and analysis.
7. When making natural resource management decisions, best available science must align with best available data. With exceptions for confidential data, the recommendation in any litigation or other public hearings must be confined to the data available at the

time. Of course, making data more readily available will greatly expand the horizons of understanding.
8. Modern techniques for data curation not only ensure proper attribution but also encourage data sharing.

Data Visualization

9. Close proximity and access to data promotes more effective data visualizations. Metadata (information about data) can convey proper data usage reliably as a proxy for direct access to the data producer.
10. The availability of cheap and open-source tools for visualization challenges the State to produce more robust, authoritative and informative data visualization tools to foster meaningful public engagement in critical environmental decisions.
11. Because data visualization often aggregates multiple data sources, data standards can help to streamline the development of visualization platforms.

Sustainable Business Models

12. Coordinated and collaborative data management must be conducted using business models that foster sustained, incremental investment and partnership with non-governmental partners.
13. There are many revenue and funding models from which to choose. A hybrid, diversified approach to funding the adopted solutions will likely protect against any single failure in the funding stream.
14. Over many years, we have seen a significant investment in agencies and organizations to conduct their data management. What is lacking is a business model to sustain the development and maintenance of data standards, integration points, web services, and data federation to facilitate synthesis across agency and issue boundaries. A sustainable, large-scale, partnership-driven infrastructure would facilitate a more comprehensive understanding of the complex California socio-environmental system.

Recommendations

Facing the challenges detailed above will require unprecedented levels of collaboration, creativity, and transparency. The solutions must build upon present investments while also disrupting the current dependency on highly

In the world of technology, individual solutions abound. This document does not directly pursue solution implementation. Rather, in advance of successful implementation, key stakeholders must first settle upon shared principles, objectives, and understanding. The collective vision in this white paper directs us toward steadier funding for technology infrastructure, more robust sharing data among agencies through a combination of incentives and mandates, and the support for an engaged and innovative technical staff.

centralized systems and processes if the State is to foster ambitious, agile technology innovation. These solutions will reside not among an exclusive cadre of insiders but at the broader intersection of all interested parties including the public, agencies, local governments, NGOs, and tribes.

To help organize the anticipated effort, we have organized our recommendations according to a schedule of near-term, mid-term, and long-term actions. Where possible, we have also indicated the expected duration of the recommended activity.

Near-term

A comprehensive *data federation strategy* should be adopted by the State to bring data together into a virtualized unity, while still preserving the autonomy of individual data repositories. Individual data systems will continue to evolve in alignment with their own individual mandates and stakeholder needs, but in addition, they must be enhanced to offer integration options for inclusion into a statewide, interagency, federated system. Such a federated system will result in a holistic understanding of the State's ecosystems while accelerating analysis and discovery for each individual member system. Implementation can be accomplished in an incremental fashion to allay concerns from data managers and address substantial decision points. The tasks ahead call for the empowerment of one or more broad-based, collaborative, interagency workgroups to achieve the following implementation-related goals:

1. Data governance oversight (p.27)
    a. review available interoperability standards for environmental data,
    b. document common metadata standards (or set of standards),
    c. seek strategic alliance with national and global initiatives that can contribute tools and web services.
        i. explore what web services / integration points exist and what needs to be developed to facilitate sharing of data.

    Duration of engagement: 1 year

2. Developing a data federation strategy with a specific, time-bound roadmap. This effort must complement the work of the data standards implementation. (p.27, 42)

    Duration of engagement: 2 years

3. Develop a business case and adopt a funding strategy in service of a sustainable business model, optimizing cost-benefit for the public good. The funding strategy and business case, once shared with strategic partners, will inspire the collaboration and cooperation necessary to motivate further efforts. (p. 41)

Duration of engagement: 2 years

Points of information:

- Whether these workgroups are singular or multiple depends largely on institutional capacity, scheduling matters, and jurisdictional concerns.
- Such efforts are currently underway. For instance, the Data Management Workgroup and the Wetland Monitoring Workgroup, both associated with the California Water Quality Monitoring Council, are conducting inventories of the State's metadata and data standards for select environmental data. Work such as this must continue, in whatever form is appropriate, to collect standards used by all of California's high-priority environmental data.
- Furthermore, data integration projects, such as those pursued by the Strategic Growth Council and Delta Restoration Network, should be encouraged as learning opportunities. Lessons gleaned from these pilots should in turn inform the data federation strategy.
- Regarding data standards, they should be promoted but not at the expense of data repository heterogeneity. Such heterogeneity enhances security and guards against the possibility of a total shutdown under a cyber-attack. Data federation preserves data heterogeneity while still advancing the dynamic sharing of data.

Mid-term

4. Embrace data of differing quality, resolution, and sources, provided that these attributes are documented according to standards. (p.27)

   Duration of engagement: ongoing

5. The State should devise a strategy for cultivating a common set of visualization tools. By leveraging talents across agency boundaries, the State can develop a knowledge-base and common set of technology libraries for data visualization development. This can decrease expenses while fostering modeling efforts, outreach support, and management engagement for more effective decision-making. (p.35)

   Duration of engagement: 2 years

6. Adopt open-source software experimentally where appropriate. (p.24, 42)
   a. A mix of open-source and proprietary solutions and tailored web services can meet ongoing needs while addressing emerging demands. Increasingly, technology must be easily upgradeable and versatile. A hybrid mix would lend stability and flexibility while also encouraging cost-effective innovation.
   b. The State must continue to recruit and retain the best and brightest technologists. For software developers and technology support staff, open-

source software holds the greatest promise for career advancement, knowledge enrichment, and solution development. The State must cultivate opportunities to employ open-source software through training and challenging career tracks for these critical positions.

Duration of engagement: ongoing

7. Investigate opportunities for supercomputer, cloud computing, and massive distributed computing projects. Initiatives led by national programs are spearheading several new systems. Investigations of California's water challenges could be accelerated or enhanced by partnering with this massive computational and data storage capability. (p.44)

   Duration of engagement: ongoing

Long-term

8. Develop and implement data management plans for all data acquired that clearly incentivize data-sharing. California should tie future funding opportunities to data transparency, similar to the National Institutes of Health and the National Science Foundation's present policies, such that the requisite time to post data are clearly defined. Recognizing that some data must necessarily be restricted at least for specified time (for example due to litigation or implications for sensitive or endangered resources), data-sharing policies should be clearly articulated with reference to state and federal laws as appropriate. California must strategically position its data management plans toward national and international initiatives and standards. Consulting contracts related to data generation should also be subject to these guidelines. (p.46)

   Duration of engagement: ongoing

These recommendations build upon the State's existing infrastructure and nascent initiatives while also offering necessary opportunities for growth at a time when our natural resource management requires well-informed and timely decisions. More than ever before, we must work together across jurisdictions and disciplinary boundaries, for our success will be measured by our collective advancement.

Enhancing the Vision for Managing California's Environmental Information

Introduction

You can't do your science without sharing your data in the geosciences...NSF [The National Science Foundation] has always had a data-sharing policy. If you create data using NSF funds, then you must share them broadly...and promptly.

–Jennifer Schopf, Director of International Networking, Indiana University

To maintain market power, to provide your staff with new growth opportunities, to build heterogeneous systems, to lower license liabilities, to be ready to scale: for all these reasons, it makes sense to have open source as an option...Integrating open source into your current systems is not revolutionary, but evolutionary, a little bit at a time.

–Paul Ramsay, Vice President - Geospatial Architect, Boundless

California's Environmental Data Summit, the product of a multi-agency request in the Delta Science Plan, was convened in June 2014 with support from state, local, federal, and private partners. The Summit challenged its attendees to embrace new ways of thinking about sharing data crucial to natural resource management. Technologists, policy-makers, scientists, and agency staffers gathered to share new approaches to persistent technical and procedural problems rendered all the more acute by the urgencies of the moment. During a period of an ever-deepening drought, compounded by pressing matters such as climate change and an aging water infrastructure, these crises force all Californians to recognize the fragile balance between our natural resources and their competing uses across the State. Under these circumstances, poor quality data and good but sequestered science can degrade mere frustration into irrevocably altered or lost ecosystems. Natural resource decisions must be based on science that both measures the rate of environmental change and informs the requisite fast pace of adaptation efforts.

"With the world's most advanced technology resources located here in California, how do we apply our intellectual resources to this problem, to ensure that sharing information becomes the norm for natural resource management rather than the exception? And how do we best complement and build upon on our past technology investments?"

The speakers and workshop sessions expressed a diversity of opinion but also shared a common drive to improve the condition of California's technology "ecosystem." Summit attendees heard from Jennifer Schopf, whose robust data sharing under the National Science Foundation's strong mandates served as exemplary guidance and inspiration. Listeners also heard from Paul Ramsay, an open source advocate and geospatial technology innovator, who contended, "Open source software and open-source culture are the new normal. They are not going anywhere." He suggested that technology managers embrace experiments with open-source software to improve outcomes, expand possibilities, and retain valuable talent. Later, during workshops, leaders in environmental technology and science shared their challenges and collectively proposed solutions. The ideas detailed in the paper below are emergent from these and subsequent discussions.

In essence, this vision document seeks to chart a course toward a more evolved data stewardship strategy, broader uses of data visualization, and more sustainable business models to foster new and productive relationships across all sectors. This vision should serve as a resource to policy makers, agency leadership, data managers, and others who wish to foster more robust and coordinated information management efforts. With the world's most advanced technology resources located here in California, how do we apply our intellectual resources to this problem, to ensure that sharing information becomes the norm for natural resource management rather than the exception? How do we best complement and build upon on our past technology investments? How do our biggest strides exemplify the most forward-thinking and effective data visualizations that convey challenging ideas comprehensibly?

Our work is to establish a vision that can be broadly shared among agency, NGO, tribal, academic, and public stakeholders. By fostering principled consensus, we wish to demonstrate that California has a broad plan for addressing its environmental data management challenges, thereby fostering a collaborative, fundable package of solutions that can be implemented over a sustained period of time with minimal disruption to established processes.

To be effective in addressing these questions, agencies, organizations, and public interests must accelerate their adoption of innovative technologies. They must share the information and data produced through these innovations across conventional jurisdictions and agency boundaries using modern, rather than dated, definitions of data sharing. And while forming partnerships with non-governmental organizations and the private sector, the public sector must adopt business models that foster sustainable technology solutions just as we seek to manage our natural resources more sustainably. In short, the need for transparency and sharing of data demands an open community of scientific information. Decision makers, analysts, and public interests collectively recognize the need for such a sharing initiative.

California's Department of Water Resources, for instance, cites the development of the California Water Plan as an example of a challenging process made more arduous by the need to acquire data from various local, state, and federal data sources. These information pathways are fraught with obstacles, varying data standards, and incompatible data access systems. Spreadsheets are circulated with calculations performed manually, producing unnecessary cost, opaque processes, and additional risk of error. Such examples are common, even when accessing data held within the State's own technology purview.

In its review of the sustainability of water and environmental management in the California Bay-Delta, the National Research Council (NRC) found that "only a synthetic, integrated, analytical approach to understanding the effects of suites of environmental factors (stressors) on the ecosystem and its components is likely to provide important insights that can lead to enhancement of the Delta and its species" (National Research Council 2012). Currently, outstanding research is being performed to address specific issues that are aligned to individual agency missions and legal mandates at specific geographic locations. Meanwhile, emerging sensor technologies and increased monitoring requirements have resulted in a deluge of data in the past decade that is often inaccessible in a timely fashion to scientists addressing related issues. In order to break out of this "silo-science" and foster the types of synthesis activities deemed essential by the National Research Council to accelerate knowledge discovery and better inform management actions and policy, it is essential that we know what data are being collected, and information regarding the quality and source. There are numerous laudable pilot efforts underway that integrate multiple sources of data as well established databases. These efforts need to be nurtured and grown into an overall strategy to share data and provide the resources necessary to maintain and protect data resources. Failure to prepare California for managing these ever-multiplying data-streams will perpetuate the disparate "silos of science," incomplete information being used in management decisions, the lack of synthesis necessary to understand landscape-scale and population-level responses, as well as the enablement of "combat" science where different groups use different data and information to justify different conclusions. Shared and accessible data with a clear understanding of the accuracy and content of the information is paramount to minimizing lawsuits and conflicts, advancing nimble management, and deepening public engagement.

At the federal level, several active initiatives and mandates address the recognized need for greater data sharing, transparency, and coordination. For instance, President Obama's Executive Order from 2013 requiring federal agencies to implement the President's Open Data Policy requires that, whenever possible, "information resources [be made] accessible, discoverable, and usable by the public [to] help fuel entrepreneurship, innovation, and scientific discovery" (Executive Office of the President, Office of Management and Budget 2013). Pursuant to this goal, the Open Water Data Initiative, conducted by the U.S. Geological Survey, represents a multiparty effort to "establish more effective working relationships and foster collaboration with state and local agencies, Indian tribes, and the private sector" with the ultimate goal of offering "advice on efforts to operate a cost effective national network of water data collection and analysis..." (ACWI 2014). Though these federally driven efforts are already

well underway, the State of California can yet seize the opportunity to influence the open data movement and, at the same time, realize its own ambitions.

These directives dovetail with other demands for action. The Governor's *California Water Action Plan* significantly calls for federal, state, tribal, and partner collaboration to fulfill the goals related to water resource "reliability, restoration and resilience": "Better technology can result in improved coordination and more accurate data for decision making" (*California Water Action Plan* 17). By tying good decision making to better technology, the Plan amplifies the urgent need for a stronger, 21st-century information management foundation.

> In the world of technology, individual solutions abound. This document does not directly pursue solution implementation. Rather, in advance of successful implementation, key stakeholders must first settle upon shared principles, objectives, and understanding. The collective vision in this white paper directs us toward steadier funding for technology infrastructure, more robust data-sharing practices among agencies through a combination of incentives and mandates, and the support for an engaged and innovative technical staff.

This white paper describes examples, practices, and recommendations that focus on California's Delta as an ideal example for yielding meaningful results in the face of pressing challenges. After all, the Delta represents a complex terrain of multiple, overlapping agency jurisdictions and discrete technological silos. Once proven in the Delta, however, this paper's recommended innovations would conceivably be applied across the State in subsequent phases.

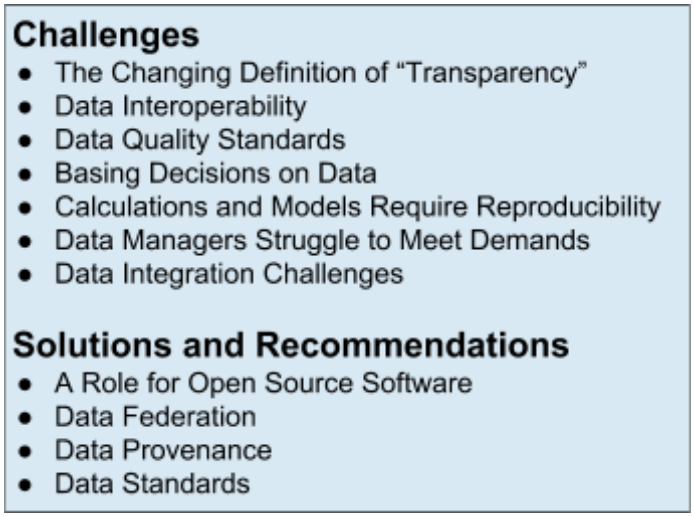
Organized into three sections, this white paper begins with a description of data-stewardship challenges and recommendations to address them. With "**data stewardship**," one would associate data governance, standards, provenance, and new models such as data federation. Then, the second section seeks to illustrate an output from more transparent, consistently organized and documented data: **data visualization**. We discuss visualization, not to the exclusion of data mining and other analytical tools, but rather as a way to foster an understanding of the tremendous value yielded by a broadly understood category of technology solutions. Finally, we close with some attention to the State's **business model** for supporting technology solutions and potential ways to span process and funding shortfalls.

Even as obstacles continue to impede us, opportunities abound. The attendees left the conference humbled by the many challenges but inspired by the ideas and the urgent needs to act. Our vision document writers -- Shakoora Azimi-Gaylon, Stephanie Fong, Peter Goodwin, Tony Hale, George Isaac, Amye Osti, Fraser Shilling, Tad Slawecki, Steven Steinberg, Mark Tompkins, Laci Videmski -- capture in their writing the most salient of the emergent ideas shared during the event. The result is a document that characterizes the many opportunities for

contemporary data management, envisions how we might identify the obstacles before us, and charts a path toward an open community of information sharing.

Evolving Data Stewardship

Data stewardship is the management and oversight of data assets to help data users discover, access, supply, and use data of known quality in a consistent manner. The importance of effective data stewardship has been examined at length at the state and national level and has been recognized for its critical value by numerous government agencies and private organizations for many decades (Ghosh 2011). While there remains general agreement that sustained data preservation and effective data access require thought and resources to manage effectively, the very meaning of data stewardship in today's world of big data and dynamic data sharing compels us to re-examine the traditional data stewardship programs. Data stewardship, conceived as a program, must continue to encompass all activities that preserve and improve information content, its accessibility and usability, while also remaining nimble to accommodate today's changing needs.



An active data stewardship program will help the environmental data providers tackle the difficult tasks of agreeing upon consistency across multiple mandates and requirements, yielding data that can be transformed into actionable information. Various groups are often committed to their own proprietary business rules and definitions. A unifying data stewardship program is necessary to work closely with all interest groups to develop and embrace common business rules and definitions that promote clear cross-communication, yet allow sufficient flexibility to encourage inclusiveness.

Sharing data allows scientists and researchers to tackle complex problems that may not have been feasible to address in the past, and encourages collaborations that blend disciplinary research so that these challenges may benefit from new approaches and new ideas.

Data governance, an essential component of a data stewardship program, is a system of decision rights and accountability for information-related processes, executed according to agreed-upon practices which describe who can take defined actions with defined information, and when—under defined circumstances—using defined and approved methods.

One of the biggest historical problems with data governance is the lack of well-defined policies that may not have established the underlying organizational structure to make it useful. The solution to this problem requires two steps:

1. The definition of the management structure to oversee the execution of the governance framework and an incentive model that rewards that execution. A data governance framework must support the needs of all the participants across all data providers, both from the perspective of the data provider and that of the data consumer.
2. A workgroup must be convened to establish best practices and coordinate technical approaches to ensure precision and completeness of data and information. An effective framework will benefit from a wide range of participation within a data governance oversight group, while all interested parties can participate in the role of data stewards.

In a traditional organizational data management environment, business rules encompassed within data controls can be used to govern both the creation and the consumption of data across the data providers. Increasingly, however, as big data analytics applications are absorbing massive data sets from external sources, the points of data generation are further and further removed from their various points of data usage. As a result, the ability to control the full data lifecycle -- from collection to processing to dissemination to usage -- must give way to a different kind of data governance.

It is important to recognize that datasets created for some functional purpose within an organization will be reused in different contexts for other purposes. This should become the expectation and managed accordingly. The implication is that data quality can no longer be measured in terms of “fitness for purpose,” but instead must be evaluated in terms of “fitness for a plurality of purposes,” taking all additional uses and quality requirements into account.

When focused internally, data governance not only enables a degree of control for data created and shared within an organization, but it empowers the data stewards to take corrective actions, either through communication with the original data owners, or by direct data intervention (that is, “correcting bad data”) when necessary.

Challenges

The technology and infrastructure necessary to make scientific data discoverable, accessible and available on demand has evolved tremendously in the last decade. With widespread access to the internet, many organizations now make scientific data available on their organizational website, or through data portals maintained and managed by government agencies, university libraries, non-profit organization or other venues. This tendency to share more broadly reflects growing interest in the data used to manage our natural resources. Nonetheless, a variety of obstacles remain which make it difficult to organize, locate and access scientific data from the many organizations that collect these data.

The Changing Definition of "Transparency"

In the past, the concept of data transparency was largely defined by the availability of datasets in *any* format. This often meant data were accessed by contacting the individual scientist or organization directly to request a copy. Data may have been delivered as a physical report or data file mailed to the requestor, and could take a substantial amount of time and energy to identify, locate and obtain. Furthermore, in many cases, such datasets would be delivered in processed formats such as summary statistics or analytical reports. These approaches to data transparency were functional from the perspective of providing an opportunity to review and assess the analysis of an individual dataset or study. However, this approach provides a limited ability to use the data for other purposes or to assess the integrity of the original, unprocessed evidence.

> "The days of mailing a postcard or making a phone call to request a scientific report are long gone. Transparency is now about getting exactly the data desired in a consistent and comparable format, with quality metadata as soon as the data are publically released."

As organizations moved into the Internet age, these reports and data files were instead posted to a website for public access. However, the underlying format of these data did not evolve significantly. Many organizations took an approach of simply posting these same reports as word processor or portable document files (PDF) in lieu of providing a physical copy. Similarly, a dataset may have been provided as a compressed ZIP file available to download in conjunction with a report and perhaps some metadata. Fundamentally, the early conversion to the web consisted of replicating previous approaches in a digital format.

As the web evolved in the mid-to-late 1990's, more organizations began to leverage the capability of web-accessible database systems. Such systems allowed for data to be posted in a native format, potentially providing a variety of tools to query the data using a web browser. Tabular data were the first to be made available in this manner, and later, spatial data using a map interface as the query tool allowed for data to be selected by location. In either scenario, data potentially started to become more useful in that specific information was now available in a digital format which could be assessed by the individual selecting and downloading it from the source, as opposed to being limited by the decisions made by the originating scientist or organization. Access to the underlying data from a given study provided a new level of transparency. However, even still, it was largely lacking in the interoperability and comparability to allow for the integration of data from multiple sources and studies. Obtaining data typically required visiting numerous websites to find relevant (and often irrelevant) data sources.

The initial attempt to remedy this issue was that of data aggregation into common data systems. Data portals were developed to bring together sources of information from multiple studies within or across organizations using a common database. These large data systems when done

well, provided new benefits of data that were discoverable in a single location and comparable across sources and studies. Because data were aggregated in a single database system, there were often issues of infrastructure and management of these large data systems that were likely to become somewhat unwieldy, difficult to manage and expensive. Data no longer remained with the original source, but rather with the system.

More recently, the application of a web-services model has begun to replace the central data portal concept. Web services are based on a set of protocols and metadata standards which allow for the cataloging and query of data from multiple participating organizations via a single portal. However, the actual data remain with the authoritative source, the organization or individual responsible for these data.

Each of these evolving solutions help to make data available, transparent, and more accessible to organizations, and all of these approaches remain in use to varying degrees. However, as the scientific community and the general public have become more familiar with the vast array of information available via the Internet, the expectation of access has evolved as well. As a general rule, people expect data to be available immediately and on demand. The days of mailing a postcard or making a phone call to request a scientific report are long gone. Transparency is now about getting exactly the data desired in a consistent and comparable format, with quality metadata as soon as the data are publically released. In some cases, even the concept of public release is blurred with some data being made available in preliminary format before quality assurance and quality control processes (QA/QC) have been performed and prior to completion of all anticipated analysis and reports.

Data Interoperability

Interoperability has emerged as an important vehicle for transparency through the proliferation of web services. Web services provide the technical means to make scientific data available, but there remains a significant degree of human effort to develop the required underlying data and metadata structures. For new and future data, extraordinary opportunities exist to plan for a level of data discovery and availability not previously possible. Accomplishing this same task for existing and historical data is a much more difficult task. However, for the analysis of long-term trends and longitudinal studies, the effort may be justified for a number of high value datasets.

Data Quality Standards

Good science and good decision support demands access to timely data of known and documented quality. To achieve analytical results in an effort to answer scientific questions, it is essential to draw upon relevant and appropriate data. When these data are sourced from existing data sources, such as those provided via web services, one’s ability to coordinate and control the data collection process is limited. Well-documented data with adequate metadata are

necessary to evaluate whether they meet the particular requirements of an analysis or as parameters for a model.

Basing Decisions on Data

When making natural resource management decisions, best available science must align with best available data. With exceptions for confidential data, the recommendation in any litigation or other public hearings must be confined to the data available at the time. Of course, making data more readily available will greatly expand the horizons of understanding. One way to achieve greater availability is through the use of “web services.”

While web services provide timely access to available data, they are not typically effective in communicating the complete universe of data. A strength of web services is that they provide the technical infrastructure necessary to make data available as soon as they are published by the originating organization. However, data publication does not always occur in a timely manner. Depending on the data type, it is often necessary to subject data to a variety of internal reviews (e.g. QA/QC, verification that data are properly structured and documented for publication, and in many cases that internal uses and/or analysis have been completed and published prior to data release). These issues may delay the release of data by days to months depending on the context. Such delays limit the ability for others to effectively utilize these data in their own decision support contexts in addition to their value in providing transparency of process and validation of results by others.

Although web-services can make data available almost immediately upon release, there is a risk that end-users of such services will not be aware that the data they are receiving through the service may not represent the complete universe of data available for a particular system location. However, web-services which also communicate information regarding the nature and content of yet-to-be-released data have a potential to buffer against misunderstanding and/or misrepresentation of results based on incomplete information.

Beyond access to data, a number of data characteristics must be assessed in selecting data for analytical and decision-making purposes. When examining the key characteristics of big data analytics, in support of science and decision-making, the analogies with the conventional approaches to data quality and data governance have several levels of data usability and are measured based on the idea of “data quality dimensions,” such as:

- Accuracy, referring to the degree to which the data values are correct;
- Completeness, which specifies the data elements that must have values;
- Consistency of related data values across different data instances;
- Sensitivity and uncertainty of the measurements recorded;
- Precision to indicate the proximity to all possible data interpretations;
- Timeliness of the data, especially as real-time data become even more broadly used.

Such measurements are generally intended to validate data using defined rules and identify errors when the input does not conform to those rules. This approach typically targets moderately-sized datasets, from known sources, with structured data, and a relatively small set of rules.

A difficult issue in big data analytics is the question of consistency. When datasets are created internally and all additional data users recognize a potential error, that issue can be communicated to the originating system's owners. The owners then have the opportunity to find the root cause of the problems and then correct the processes that led to the errors. With big data systems that absorb massive volumes of data originating externally, the tools can yield insights and identify inconsistencies in the data. This is one of the strengths of the big data tools. However, there are limited opportunities to engage process owners to influence modifications to the source. If, as a big data user, you opt to "correct" the potential data flaw unilaterally, you may be introducing an inconsistency with the original source, which at worst can lead to incorrect conclusions and flawed decision-making.

Calculations and Models Require Reproducibility

Data systems which make it easy for users to identify the accuracy, consistency and completeness of data are essential when models are utilized in analysis and planning. A keystone of scientific inquiry is reproducibility. The ability for other scientists to replicate results requires not only that analytical methods and models are well documented, but also that the selection of the data used in the analysis is also documented and reproducible.

While a web-services approach to data discovery and access provides many benefits, there is a significant challenge when such services are used to source data. Because web-services may provide access to live databases, the data identified and acquired at the time of a study may change at some future time when another scientist seeks to reproduce the analysis to validate the results. Therefore an important consideration is that of capturing the information or metadata necessary to replicate the data selection process.

As more such analytical tools are linked to web-services, providing an ability to access data "on-the-fly" there is need for development of processes for tracking of data lineage or provenance. In this way, if data coming from the web have changed since the original analysis, there is a means to either 1) replicate the data acquisition request as of the time of the original analysis, or 2) identify and assess any changes to the source data due to updates (corrections, additions, or deletions) to the underlying data.

Data Managers Struggle to Meet Demands

Many hard-working data managers/stewards lack clear directives, common standards, and technology resources to meet the challenges and expectations of today's data users. While the expectations for integrated, web-based data services has become common, the resources, direction and policies for implementation have not kept pace. In a world where there is a perception that almost anything can be located, researched and acquired and delivered instantly (online) or within a day or two, there is a perception that agencies and organizations *should* be capable of providing similar capabilities to their constituencies as well.

Two key features must be addressed to improve this capability for scientific data:

1. Data integration, to ensure that data are discoverable, documented and accessible.
2. Institutional capacity, to provide the human capital, technological infrastructure and policies and procedures necessary to support accessible data systems.

Data Integration Challenges

Among the most fundamental challenges in the process of data integration is setting realistic expectations. The term "data integration" suggests a perfect coordination of diversified databases, software, equipment, and personnel into a smoothly functioning alliance using comprehensive systems of information management. Yet the work ahead is difficult. We must recognize that formidable integration challenges remain, even as we embrace promising new processes and big data integration tools.

Heterogeneous Data

For some users, data integration involves synchronizing huge quantities of variable, heterogeneous data resulting from internal legacy systems that vary in data format. Legacy systems may have been created around flat file, network, or hierarchical databases, unlike newer generations of databases which use relational data structures. Data in different formats from external sources continue to be added to the legacy databases to improve the value of the information. Each generation, product, and home-grown system has unique demands to fulfill in order to store or extract data. Data integration can involve various strategies for coping with heterogeneity. In some cases, the effort becomes a major exercise in data homogenization, which may not enhance the quality of the data offered.

Bad Data

Data quality is a top concern in any data integration strategy. Legacy data must be cleaned up prior to conversion and integration, to avoid serious data problems later. It is not unusual for undiscovered data quality problems to emerge in the process of cleaning information for use by the integrated system. The issue of bad data leads to procedures for regularly auditing the quality of information used. (However, it is important to note, while bad data is anathema to good decisions, variable uncertainty and precision are features of scientific data and can be documented and accommodated.)

Unanticipated Costs

Data integration costs are fueled largely by items that are difficult to quantify and thus predict. It is important to note that, regardless of efforts to streamline maintenance, the realities of a fully functioning data integration system may demand a great deal more maintenance than could be anticipated.

Lack of Cooperation and Coordination

User groups within an agency may have developed databases on their own, sometimes independently from others that are highly responsive to the users' particular needs. It is natural that owners of these functioning standalone units might be skeptical that the new system would support their needs as effectively.

One entity may not want the data they collect and track to be at all times transparently visible to others without the opportunity to address the nuances of what the data appear to demonstrate to others. Owners or users may fear that others without appreciation of the data's importance might gain more control over how data is managed and accessed regionally.

Recommendations

There is no single panacea for California's data stewardship challenges, but there are steps we can take -- some incremental, some boldly transformative -- to remedy many of the woes described above.

A Role for Open Source Software

Data management for California's environmental resources will require the effective leveraging of the full continuum of software from completely open source to proprietary. Generally, in public agencies, open source is shunned, but there are opportunities to marry open-source and proprietary solutions together into a productive whole.

Open source software and forms of proprietary software each have a role in providing reproducibility and accessibility to the general public at little to no cost, provided that they offer sustainable technology frameworks. An open source software license refers to software distributed without charge, and in many cases with the option to modify or customize the underlying source code. By contrast, proprietary software requires the end user to purchase a license and provides limited to no ability for the end-user to modify it for their purposes. Both of these software distribution models have proven successful in a variety of contexts, including the development of data systems designed for the discovery, exploration and access to scientific data. The choice of software platforms is one which raises a number of key questions in context of developing and sharing of data. Most crucial is the ability of the organization to install, maintain and support the selected software platforms.

Commercially acquired software from large vendors has the advantage of providing an “out-of-the-box” solution, complete with technical support, documentation and training opportunities. This lends itself to getting systems up and running more quickly, and the added comfort of having someone to call should problems arise. However, these same software platforms limit one’s ability to add new features, typically requiring the end user to wait for a new release, often accompanied by an invoice for the upgrade and/or ongoing support. Additionally, some commercial platforms may not be fully compliant with standards put forward by international groups for data interoperability. Rather, these tools may opt to use their own proprietary data formats which can lead to difficulty in achieving the interoperability desired.

Open source software, by contrast typically requires a higher level or expertise to install and maintain. In particular, they may require knowledge of particular programming languages, operating systems or hardware platforms. Support is obtained from user communities rather than a helpdesk. However, these differences also provide the ability to customize the software, to become part of a larger community of motivated end-users and to save the ongoing costs for upgrades and support. Many of the more popular open-source products have generated 3rd party support and training that rivals (and sometimes exceeds) opportunities related to commercial products. Many open source tools are standards-based which helps to ensure interoperability.

While there is no single “correct” choice between open source and proprietary software, it is essential that all software decisions be considered carefully.

1. What is the organizational ability to support the selected software (e.g. cost, expertise, upkeep)?
2. Is it standards-based (e.g. adheres to accepted data structures and communication protocols)?
3. Does it have a substantial user-base and support structure (e.g. it is sufficiently popular that it is likely to be maintained for a significant period of time into the future)?
4. How portable is the solution? Can it be ported to other environments and its scientific results replicated elsewhere?

By considering these questions carefully at the outset, there can be a reasonable expectation that the selected software solution will not become outdated and require replacement on an unreasonably short time-horizon. However, it is equally important to have a strategy in place to plan for upgrades and/or replacement of systems over time as the technology and capabilities evolve.

Data Federation

Within California, many enterprises in various sectors – both within the state government and beyond -- collect environmental data to support resource management decision-making. These enterprises have developed proven, effective and sustainable processes and workflows to ingest, store and disseminate their data for their own specialized purposes (and often find their data to be of use to others). In the case of the State, each agency will often manage its own data repository -- such as Geotracker, the California Environmental Exchange Network (CEDEN), and the Biogeographic Information and Observation System (BIOS) -- to ensure alignment between the agency's policy or regulatory mandates and its collected data. These isolated data silos can be useful, even essential, within certain constraints, but they fail to scale beyond the jurisdiction of each individual agency or organization. One significant remedy for this isolation is a technology called "data federation." Put simply, data federation facilitates the gathering of multiple data repositories into a single virtual database. The source databases remain effectively unmodified – their respective mandates and governance may continue unabated – but the aggregated virtual database can address questions that exceed the scope of any single repository. In this way, the utility and value of the individual databases can be cost-effectively leveraged by linking them into an easily accessible federated network for use by resource managers, regulators, and the general public.

Federation is not easy, but it is nevertheless very important. In such a network, standardization of data (and metadata) access leads to a heterogeneous network-of-networks. There are competing standards, and reasonable decisions must be made regarding their broader application. A federated data model supported by common standards, for instance, supports the international Group on Earth Observation (GEO) Portal, which searches metadata records harvested via standardized web services from hundreds of digital repositories.

"Integration, or rather interconnection, of California's existing data repositories into a federated data model will emphasize evolution over revolution."

Integration, or interconnection, of California's existing data repositories into a federated data model will emphasize evolution over revolution. Data owners will not be expected to throw away established, working data collection and management systems, but will instead be encouraged to add functionality that improves interoperability (making data easy to access) and discoverability (making data easy to find). Compare the limited paths for data flow in Figure 1 (depicting independent data repositories) to the wealth of options in Figure 2 (depicting a potential federated architecture).

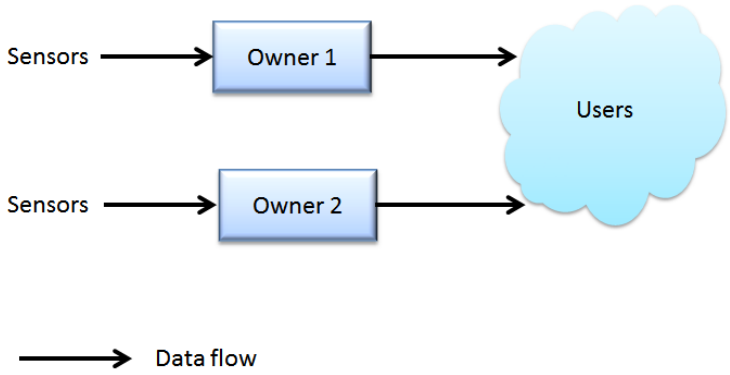


Figure 1. Depiction of data flow through independent repositories. Users take on the burdens of locating and accessing data from each repository separately.

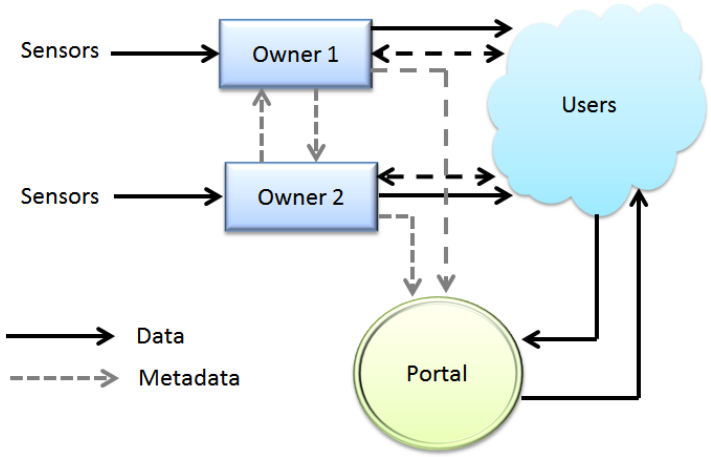


Figure 2. Depiction of data flow in an illustrative federated data network. Data owners publicize (with metadata) and publish their data, allowing other data owners and third parties to re-publicize and provide links to the original owners' data. Users are able to readily interact with owners and portals to actively discover and access data.

The recommendation of this group is to implement a federated network of networks that interconnects a significant portion of California's diverse data repositories. A well-managed Implementation might include the following major activities:

- Development of a California-wide consensus that fosters thinking beyond existing data silos and encourages collaboration between data owners.
- Identification and engagement of key data repositories and owners
- Review of holdings and practices for key data to inform standards-based recommendations for federation.
- Federation through enabling key data holders to make metadata and data available through standards-based web services – generally added to their existing data systems to leverage their long-term investments.
-

- To ensure that heterogeneous data can be properly aggregated, data providers must be involved in ongoing communication regarding data transformations and the application of requisite standards.
- All of this work must be achieved through the work of a collaborative, interagency workgroup charged with developing a data federation strategy with a specific, time-bound roadmap.

The federated network will also allow users within and beyond state agencies to discover newly exposed public data. As the reach of the network grows, data owners will be better able to identify synergies in data-collection efforts, in turn optimizing data stewardship as a whole and even identifying critical data gaps. With these promising outcomes , a data federation project must nevertheless bear in mind these important enabling strategies:

- Any large-scale data integration project, regardless of model, demands that executive management be fully on board.
- Informing and involving the diversity of players during the crucial requirements analysis stage, and then in each subsequent phase and step, is probably the single most effective way to gain buy-in, trust, and cooperation.
- Incremental education is easier to impart than after-the-fact training, particularly since it addresses both the capabilities and limitations of the system, helping to calibrate appropriate expectations along the way.
- Done correctly, such an approach will incentivize the reuse and sharing of data.

One common example for data federation output is Zillow, the online resource for exploring real estate around the country. It benefits from public data that has been available for years, but it pulls it together in an easy, accessible way. People can, in turn, query Zillow for the aggregated data. The data sources -- the individual public data repositories in various local governments -- are not forced to change their mandates. Yet Zillow can discover and present their data.

A comprehensive data federation implementation will take time. While this effort is underway, there are data integration platforms that can be employed to bridge the gap between the current state of affairs and the ideal. These tools do not depend on data federation for their effectiveness. Palintir's data analytics, for instance, can help with decision-making through the use of data gathered opportunistically for a defined management purpose. Tools such as Palintir's can play a role in helping to provide an "onramp" to the federation solutions described in this paper.

Data Provenance

Data provenance -- the measures used to track sources of data, their transformations, and approved uses -- will prove essential to the successful adoption of federated data sharing. When data are disassociated from their points of origin, we must take extra steps to ensure proper attribution. Which organization or scientist collected these data? Standardized metadata

development practices can address this question. Moreover, data provenance also accounts for the transformations of data during processing and analysis. What datasets contributed to this analysis? Who modified these data as they were transformed from a “raw” source to one that was checked for quality assurance? Even data usage instructions, indicating the interpretive limits of the data, fall under the rubric of data provenance.

These families of metadata solutions are all important insofar as they mollify concerns over inadvertent expropriation of data from their rightful place of origin. Data providers will feel more comfortable yielding data to a federated system if reassured that the data will be properly attributed, understood, and used. Furthermore, regular citations using provenance systems help data collectors and data repositories to demonstrate the value of their activities. Within a federation, provenance can also provide transparency and scientific credibility to the use of datasets by documenting analytical transformations in the development of a decision-support product.

The recommendation of this group is to promote a culture of data accountability and transparency through provenance and the attendant development of appropriate practices and tools to make implementation of provenance straightforward and easy. Specific measures may include:

- Development of a California-wide consensus that publication of data should, at least within a certain regulatory and legal setting, require publication of data provenance.
- Identification of both minimum and desirable levels of information to include about data sources and data paths to establish provenance to a satisfactory degree.
- Documentation of existing tools, such as the California Data Library’s DOI generator, that can be used to identify data sources.
- Documentation and/or development of tools that make it feasible for users to track data’s “chain of custody”.

Making provenance easy to track for users of California’s environmental datasets will improve the reliability and credibility of management decisions while reinforcing appreciation of the value of data collection and data stewards.

Data Standards

The federation of California’s data repositories will rely heavily on participants’ commitment to standards-based metadata publication and data delivery. Consistent implementation of standards will lead to a robust, well-interconnected data network. These standards must promote not only clarity and consistency on data types and formats but also data quality. In this way, data standards use metadata as a vehicle to ensure that data users and data processors are aware of any limitations related to data quality or usage. In a properly operating system, everyone is incentivized to maintain metadata. However, deciding upon data and metadata standards for the diverse range of agencies, stakeholders, and data types will be no small task.

We recommend surveying the State's data stewards for information regarding standards currently in use. Then the hard work of adopting and incentivizing data standards should prefigure a broader federation effort. The Open Geospatial Consortium (OGC) suite of standards provides a useful starting point for discussion of existing standards applicable to the California environmental data enterprise as well as of mechanisms for addressing new needs. The OGC's consensus-based, transparent efforts have led to the promulgation of a wide range of standards that promote interoperability (e.g. Water ML, Sensor Observation Service) and discoverability (e.g. Catalog Service for Web), some of which are already in use by California data owners. These standards have been adopted in both proprietary and open-source software platforms, offering a high degree of flexibility in application to various technology platforms (e.g. GeoNetwork and ESRI GeoPortal for metadata). The OGC is also an excellent resource should new directions in standards be necessary; we recommend that:

- Representatives of the new federated network participate in the OGC Interoperability Program to provide access and direction to a worldwide community focused on improving interoperability through web services and related technologies.
    - For California, explore what web services / integration points exist and what needs to be developed to facilitate sharing of data.
- California should avail itself of interjurisdictional working groups, such as those serving the California Water Quality Monitoring Council, to ensure that challenges to implementing standards can be addressed equitably and openly.[1]
- Embrace data of differing quality, resolution, and sources, provided that these attributes are documented according to standards.

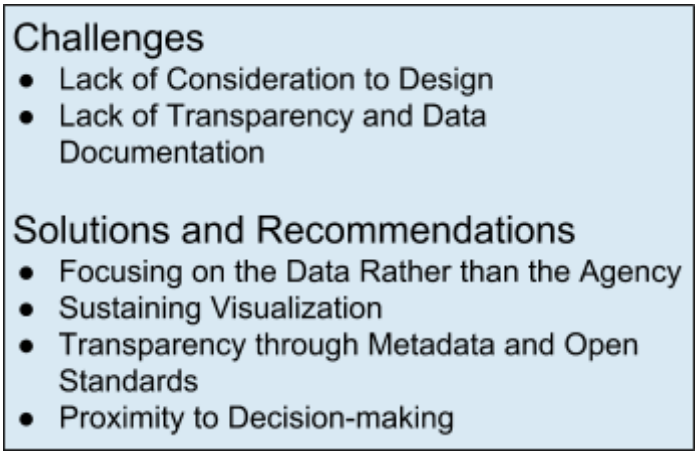
However, there will undoubtedly be data owners and datasets for which compliance will be difficult or even undesirable leading to three additional recommendations:

- Assistance should be made available to data owners who do not have knowledge or resources to implement standards-based metadata and data services. Assistance can range from training and mentoring of staff, to funding for on-site expert implementation, to handing off of standards-compliant services to a third party.
- Recognition must be included that certain important data content may not be readily coerced into standards-ready formats. This content may either be made available "as-is", or efforts be made to extend existing standards or develop new standards and services that meet analytical use cases.
- Data integration platforms not depending on federation should proceed on a parallel track to help ensure that such non-compliant datasets can still be integrated into broader decision-making processes.

[1] The Water Quality Monitoring Council has already made significant strides in advancing many of the goals articulated in this white paper through its portals and publications, including "Maximizing the Efficiency and Effectiveness of Water Quality Data Collection and Dissemination" (2008) and "A Comprehensive Monitoring Program Strategy for California" (2010).

Data Visualization

The vision described in this white paper concerns the management of "information" -- not just data. We might think of data as the raw material from which one might derive *information*: data that is analyzed, contextualized, and interpreted by experts. Paul R. Gamble and John Blackwell write, "Classically, information is defined as data that are endowed with meaning and purpose." When that information reaches its intended audience, a transaction is complete. The information is finally transformed into knowledge.



Data visualization is merely one example of several tools that help data managers, scientists, researchers, policy makers, educators, natural resource managers and others transform data into information and knowledge. In this section, we could have also described data-mining, knowledge-discovery and decision-support tools that help data users interact with large data stores. Or we could have detailed the important work that sophisticated analytical tools perform in interpreting heterogeneous data, thereby informing timely decisions. However, for the sake of clarity and concision, we are using data visualization as a metonym to stand for the many forms of data interpretation that would depend heavily on the improved data stewardship practices described in the preceding section. Because it is broadly understandable and accessible, data visualization is highly in demand, and it can be fostered by virtue of stronger transparency, data standards, and greater data access.

Visualization has truly experienced growth and change in today's information-rich, big-data era. Whether we are liberating data to make them more easily accessible or integrating unlikely sources of data to achieve new insights, the techniques and tools of data visualization have evolved in surprising ways. Naturally, this growth and attention is an exciting phenomenon that has empowered many to participate and contribute, but we must remain critical of present shortcomings and promote rigorous methods among the visualization community.

To support most science-based decisions, the decision-maker and stakeholder should be able to interrogate the data that support and inform that decision. This requires that the data developer or manager display the data in a way that transparently conveys the most meaning in the simplest way to inform a decision. A variety of tools exist to accomplish this, ranging from the simplest mapping and graphing tools, to statistical modeling approaches that convey the

data and whether or not a significant trend or relationship exists. Considering the innovative new tools for visual data interpretation and the urgent need to recognize patterns in our natural resource management, data visualization, in all of its forms, becomes a crucial instrument to help advance science-based decision-making.

How is data visualization an important catalyst of change?

While data visualization is not the only way information is consumed and exposed, it is an important tool for distilling heterogeneous information into a cohesive whole and communicating discoveries to multiple audiences. When we discuss data visualization in the context of our natural resource agencies, we often think about the use of information by decision-makers, but there is an increasing demand for transparency, accountability, and responsiveness driven by a general audience - "the public." This polity is a tangible force, and one that must also be continually addressed to ensure meaningful public participation.

Defining Data Visualization

The data is a simplification--an abstraction--of the real world. So when you visualize data, you visualize an abstraction of the world, or at least some tiny facet of it. Visualization is an abstraction of data, so in the end, you end up with an abstraction of an abstraction, which creates an interesting challenge. -- Nathan Yau, Data Points

When done appropriately, visualizations allow us to prioritize certain facets of the data that provide additional value to the dataset beyond its raw state. In Figure 3, we see the product of vast quantities of data, distilled down to a simple, monochromatic comparison of two maps. They show the historical and present-day Delta as seen through the lens of marshes. What might have otherwise been lost in the details is made very clear through this visualization.

Figure 3: A Delta Transformed: Ecological Functions, Spatial Metrics, and Landscape Change in the Sacramento-San Joaquin Delta. *These two maps, placed side-by-side in greatly simplified form, dramatize the change over time in the Delta's system of interconnected marshes.*

A multidimensional dataset such as timeseries (ex. a single geographic location might have many measurements in time), when represented in a dynamic visualization, reveals behaviors and patterns that would be difficult to interpret from a spreadsheet (tabular data), or even a static visualization. When done incorrectly, however, data visualization can communicate virtual certainty ("seeing is believing") where there is actually a degree of uncertainty. This tension between communicating across domains and losing the context for data uncertainty is especially poignant with data visualization, but this tension can be managed through rigorous practices.

While static visualizations still dominate both print and web publications, we are witnessing a rapid shift on the web toward dynamic visualizations that allow for users to query and filter data in unique and imaginative ways. Today's world of increasingly sophisticated and innovative online tools and infographics are inviting data providers to offer such tools. Software products can consume and present vast quantities of data, distill them down to their essentials, and allow users to "drill down" and gain access to the raw data sources.

However, these dynamic displays come with attendant risk. Considering a medium that allows for self-guided exploration, we must remain vigilant about methods and practices that lead to

inappropriate conclusions. Similar to the world that witnessed the advent of the motion picture, the data visualization community is still trying to fully absorb the opportunities and relative constraints inherent to dynamic visualizations in this new age.

Examples of Good Data Communication

Despite the challenges, there are many great tools currently in use for communicating and analyzing data intended for sharing. State and federal agencies have worked hard to share their extensive and long-term datasets. These examples are produced with a stamp of authority due to the close proximity of the data producers to the data interpreters. Probably the largest system is the US Environmental Protection Agency's Water Quality Exchange system (WQX, http://www.epa.gov/storet/wqx/). This system uses an updated STORET database system to collect data from trusted sources and share them via mapping and querying tools. For the California Water Plan, Update 2013, the California Department of Water Resources and the US Environmental Protection Agency supported the development of water sustainability indicators in California, including a web tool to share information and data (http://indicators.ucdavis.edu/water). Finally, a few years ago, State and Federal Contractors Water Agency and Metropolitan Water District of Southern California contracted a private company, 34 North, to develop a data-sharing tool for the Bay-Delta system: Bay Delta Live (http://www.baydeltalive.com/), which provides map-based and tabular presentations of live and stored water supply and quality information for the Delta region.

Challenges

Lack of Consideration to Design

While data visualization efforts are already making great strides within the agency domain, they are not often deployed in ways that foster the greatest interaction and collaboration. The challenge of good data visualizations is ultimately a design challenge. Good design is not decoration. A well-conceived design and a keen designer draw out meaning or utility in ways that do not distract or mislead, but rather address a problem in an elegant and concise manner. Good data visualization assembles an appropriate abstraction of the data, presenting and revealing patterns inherent to the data.

It is therefore crucial that the designer be made familiar with the data to understand their limits and degree of nuance. Without readily available documentation for the State's data and metadata, data availability alone will not contribute to enhanced knowledge. Poor data visualizations are often due, in large part, to the authors not fully understanding the data that they are manipulating.

Lack of Data Transparency and Documentation

In the prior section on data stewardship, we discussed the need for data standards. Nowhere is that challenge more acute than when applied to the development of data visualization products. Without data attribute, data quality, and metadata standards, the aggregation often performed in the course of data visualizations becomes onerous, inaccurate, and cost-prohibitive. Cheap and inexpensive visualization tools are rendered more expensive by the substantial effort to ensure data fidelity and quality. And ultimately, the magnitude of insight and confidence in decision-making is dramatically reduced.

Recommendations

Focusing on the Data Rather than the Agency

Data visualization, when deployed in targeted, well-informed, and thoughtful ways, can often become a powerful democratizing force by closing knowledge gaps and focusing debates on the science. Sharing data can turn sometimes contentious discussion centered on responsible agencies to discussions that are centered on the data, which is arguably where the discussions belong.

The importance of sharing information and data is highlighted in two complementary studies of farming activity in California.

> "Some litmus tests for useful tools are that they offer data in an understandable form, provide some analytical capacity, support education and/or decision-making, and grant users broad latitude to view data in engaging, exploratory ways."

In one case (Haden et al., 2012), scientists found that Central Valley farmers were knowledgeable and concerned about local and global consequences of climate change and this knowledge resulted in changes in mitigation practices (reducing their own impacts) and adaptation practices (installing micro-drip irrigation to conserve water). In another case (Lubell et al., 2011), viticulturalists were found to be more likely to innovate with more knowledge about environmental and economic benefits of different practices. The scientists in this study suggested that resolving knowledge gaps was an important contributor to increased use of sustainable agriculture practices. Both of these cases suggest that sharing data in a form useable by broad stakeholder groups is likely to result in the behavioral changes necessary to protect valued ecosystem and social benefits.

Sustaining Visualization

Within the public agency itself, there are opportunities to make incremental changes to existing technology and workflows, which facilitate both the exposure and consumption of data. We recommend that the data publisher give careful consideration to the following aspects of data visualization:

- Pursue clarity around type of audience and their respective needs. Data visualization is keenly attached to audience, as discussed above. Ensure successful visualization through effective audience feedback and analysis.
- Foster a standards-based infrastructure (core system, API, apps) through the consistent use of web services with robust documentation. Agency data systems, despite ambitious open-government initiatives, often lack documented web services to facilitate ad-hoc data querying. This is a critical component for data users.
- Publish analytics on data usage. In turn, this information will help to prioritize effort associated with the development of future visualization tools.
- The State should devise a strategy for cultivating a common set of visualization tools. By leveraging talents across agency boundaries, the State can develop a knowledge-base and common set of technology libraries for data visualization development. This can decrease expenses while fostering modeling efforts, outreach support, and management engagement for more effective decision-making.

Transparency through Metadata and Open Standards

Besides sharing data, it is also important to share information about the data – the metadata. These metadata describe how and why data were collected, as well as other important information. Many datasets are not accompanied by metadata, because this is a more recent concern. However, metadata help understand the data provenance (the pathway that data travel to arrive at the user) and can build trust in the data from users not familiar with the data providers. The global standard for metadata standards come from the Dublin Core Metadata initiative (http://dublincore.org/), which advocates for core sets of metadata for all data types. Related to this initiative is the World Wide Web Consortium (http://www.w3.org/standards/), which has also developed metadata and other standards for web applications. The Federal Geographic Data Committee (FGDC, http://www.fgdc.gov/metadata/geospatial-metadata-standards) has developed standards for metadata for spatial data.

Furthermore, to foster consistent openness and access to derivative data products, the State might also embrace the movement advocating “copyleft.” A play on the customary term “copyright,” “copyleft” does not seek to protect a creator’s ownership rights, but rather seeks to ensure that data and products placed in the public domain remain non-proprietary even as they are absorbed into derivative products, such as visualization tools.[2] This still nascent form of

[2] https://www.gnu.org/copyleft/

licensing might be considered by State decision-makers, realizing that its restrictions would promote transparency at the cost of potential commercial enterprise.

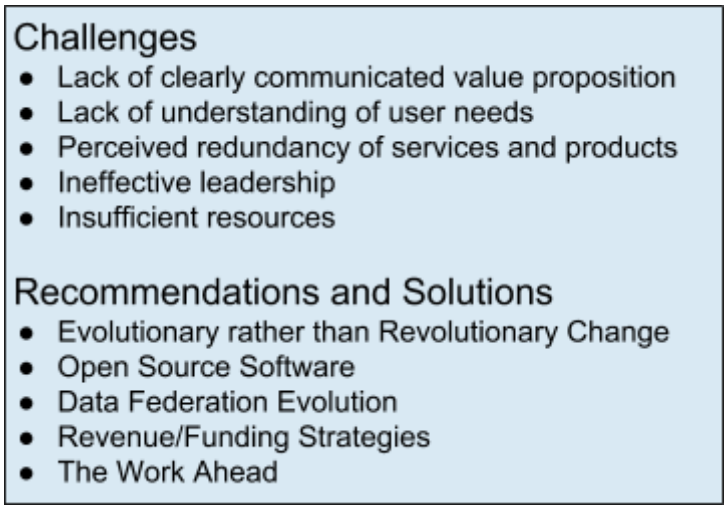
Proximity to Decision-making

Communicating data requires conscious decisions about what data to show and how. These decisions are ultimately informed by what kind of decision or process is being served by visualizing the data. Communicating data and about data is a an essential part of science-based decision-making in a democracy and is expected by the vast majority of stakeholders. There are many tools available to communicate data to users, so the choice of tool may not be as important as the intention of the data developer and manager. Some litmus tests for useful tools are that they offer data in an understandable form, provide some analytical capacity, support education and/or decision-making, and grant users broad latitude to view data in different, exploratory ways.

Creating a Sustainable Business Model

Empirical evidence seems to prove that companies relying more on data-driven decision-making are performing better in terms of productivity and profitability.

–McAfee and Bryjolfsson, 2012



A major theme echoed throughout the 2014 Environmental Data Summit is that environmental data are “key resources” wanted, needed, and used by key stakeholders. Reliable data are required by all levels of natural resource management including academia, science, operations and policy -- now more than ever. Ironically enough, even as the State faces an unprecedented crisis of natural resource scarcity, our State’s information resources suffer not from scarcity but from a lack of clarity, accessibility, and focus. Though the State is parched, it remains awash with data that it cannot use as effectively as it must. No one questions the collective importance and value of these data. They are undeniably critical to California’s effective resource management. Nevertheless, they often remain sequestered in agency silos that obscure a holistic view that would yield the deeper, timelier, and more insightful information demanded by today’s data consumers. The need to collaborate and integrate this data must drive our present urgency, but we must also develop a structure to sustain our efforts well beyond the immediate moment.

In prior sections, we have detailed many of the features necessary to modernize California's information management infrastructure in service of data integration, data sharing, and data visualization. In this section, we describe the challenges and opportunities in developing a *sustainable* business model to support a new era in data management -- enhancements that would, in many cases, build upon the achievements of many creative and hard-working agency staff who currently struggle to meet the demands for their information resources. How can we advance the State's business practices to avail itself of the advanced technological resources that reside in its proverbial "back yard"? How can we ensure that the State's data remains available to the many audiences who must consume it? And how do we maintain and recognize the value for the State's many critical data repositories and associated staffing resources, while we integrate the data into a more interoperable, cross-jurisdictional federation?

Attendees of the Environmental Data Summit remarked in agreement, "This is fundamentally a people problem -- not a technology problem." Indeed, the primary obstacles are organizational in nature. The technology solutions are numerous, but we lack coordination, funding support, and common goals. California's many agencies often face obstacles in working collectively to manage and advance its technology resources, to employ modern computing paradigms, and to promote transparency and accessibility. A sustainable business model can address some critical shortcomings.

A "Business Case"

The notion of a business case or business model is traditionally associated with a private entity seeking funds for a specific investment plan. This white paper applies this concept to a public investment. Traditionally, a business case details a clear business model with value propositions, market segments, and a path to market that includes how to package, market and deliver the goods and or services. As with private entities, the public agencies could benefit from clearly defining their products and identifying to whom they will be delivered. This information will, in turn, contribute to a sustainable business model. The business case is a stepping stone toward a business model.

A business model that is sustainable, in its association with technology, fosters longevity and stability of the investment even as it also encourages innovation to meet evolving needs. It establishes a framework for partnerships with multiple agencies and non-governmental organizations while still retaining a reliable core technology infrastructure. Such a model matches the latest in technology innovation not for the sake of innovation but to ensure that the State continues to fulfill the obligations of open government, that it continues to comply with the evolving definitions for accessibility, transparency, and effective data stewardship. In this sense, the term "sustainable" refers at once to the funding, the technology choices, and the processes used to ensure the technology's longevity. Just as our stakeholders' pursuit of environmental sustainability is multifaceted, so is the sustainability associated with a technology business model.

In order to serve these needs and develop a business model to support them, some fundamental steps are required:

- **Inventory Analysis**

  A comprehensive assessment (inventory analysis) of key data to support information needs of resource management components and stakeholders. This analysis will provide the business case with an understanding of existing data and its associated costs.

- **Cost-Benefit Analysis**

  A completed inventory of data and its associated attributes will provide the necessary information needed to perform a basic cost-benefit analysis.[3]

- **Market Segmentation Analysis**

  Building on the results of the inventory analysis and additional surveys, we should be able to accurately determine the market segments (audience) for the fundamental data. Some audiences may be difficult to characterize, particularly in the cases of an audience called "general public," but nevertheless, these data consumers can often be described categorically. Consider the table below, for example:

Market Segment	Quantity	Description
Government Organizations		
Agencies	30 Agencies	Regulation and Policy
Universities	50 Universities	Research
Private Individuals		
Land Managers, Farmers	800,000 acres	Management and meeting regulatory requirements
Fisherman	$23 B Industry	Water quality
Professional Services		
Modelers	$30 M industry	Forecasting, implementation
Application Development	$500 M Industry	Decision support tools, apps

Figure 4: Market segmentation concepts can be applied to public data products.

By identifying the audience for environmental data in this way, market segmentation will help characterize industry willingness to pay and help to develop targeted funding plans.

Without this business case development, the necessary infrastructure investments that demand a diversity of stakeholders and a diversity of funding sources may continue to be as elusive.

[3] After step 1, grouping data into categories and making analysis based on a category of data may be more efficient.

Challenges

Given the diversity of data types, users, and needs in the California natural resources management sector, there will no doubt be a wide array of significant obstacles with the potential to thwart the establishment of a sustainable business model for providing easy and open access to natural resources data generated in the State. Listed below are the five most important obstacles that must be addressed immediately, recursively, and continuously to ensure the sustainability of this effort.

1. Lack of Clearly Communicated Value Proposition

The biggest obstacle to the sustainability of this business model will be an inability to provide value continuously and reliably to the community of users it must serve. This will be difficult to do because natural resources management has such a wide variety of data types, workflows, existing management structures, and management concerns. If we do not develop a strong business case that can clearly quantify the value offered by the data and its applications to problem-solving, then the data's value is likely to be underestimated.

2. Lack of Understanding of User Needs

Closely related to the value proposition, this obstacle is a failure to understand the true needs of the California natural resources management community, which could result in the development of system architecture and tools that don't fit the workflows of those ultimately needing to use the data. A sustainable business model will address this by first completing comprehensive surveys of the needs of the community, performing fit-gap analysis, next completing pilot implementations of the data system, and finally building in mechanisms that allow users to identify and resolve system deficiencies with respect to the collective needs of the community.

3. Perceived Redundancy of Services and Products

This obstacle has already reared its head in the run-up to the 2014 Data Summit. Many stakeholders responded by saying "But we're already doing this!" without realizing the differences in scope and functionality expected from this effort. A sustainable business model will address this issue by completing a comprehensive survey of all stakeholders who feel they are in some way already doing some of the work of this effort, and then identifying overlaps and synergies so that these can be harnessed to best effect.

4. Ineffective Coordination

While there are many capable technology leaders in California, both inside and outside of government, the challenges we face require a broadly coordinated effort. Without committed and resourced leadership from a person or persons who grasp the full scope of the data uses, workflows, analytics, and problems faced in California natural resources management, this effort will fail. A sustainable business model will provide dedicated resources and powers to deliver the products of this effort.

5. Insufficient Resources

Ultimately, all businesses live or die on the balance between their costs and revenues. This obstacle is a very real danger to this effort as it will, at least initially, be viewed as a "connective" effort outside the core needs of regional or project-level data collection, analysis, and management. A sustainable business model will address this issue by rapidly integrating the products of this effort into ongoing decision-making processes and organizations such that their ability to effectively function becomes dependent on the broad access to data provided by this effort. This could be accomplished through several high profile pilot projects, and ultimately combined with the various revenue/funding models outlined below and in Appendix A.

Recommendations

The steps associated with developing a business case, articulated above, must precede a sustainable business model, but more must be done to ensure its success. To cultivate a sustainable model, we must also seed an initial investment that serves as a catalyst. It must target technology that is likely to encourage and justify further investments of time, attention, and resources. In other words, it must demonstrate its value quickly and in a collaborative fashion to incentivize further contributions.

Evolutionary Rather Than Revolutionary Change

If a fast return is key, we must also recognize the pressure to go slow in the public domain. The term "disruptive innovation" has entered common parlance these days. First coined by Harvard professor Clayton M. Christensen in 1997, this term is now used commonly in high finance and technology as a positive quality attributed to trailblazers or those who swim against the current

toward a new and better tomorrow.[4] To a certain degree, this white paper has been advocating for "disruption" of the usual way of doing business. In the context of public entities, however, disruption is not usually positive. Regularity, predictability, and process form the foundation of public agency operations. This should not surprise anyone. Disruption is risky. We should not expect public funds to be managed in risky ways. How then do we encourage bold innovation while still respecting the State's fundamental orientation?

"This is an important structural aspect to data federation: the different data repositories (nodes) are not subsumed to a central core. Rather, they contribute to a central virtual repository that bridges across heterogeneous data, making the collective whole appear seamless, but the contributing data repositories continue their work unchanged."

We recommend an approach that balances the benefits of disruptive approaches to technology with the perceived risks of disruptive approaches. For example, concerns about disruptive change should not be used as an excuse to continue investments in legacy systems when those systems no longer meet the demands of the user community. On the other hand, disruptive approaches should not be implemented haphazardly. Rather, innovation can be implemented through demonstrations of high value and low risk benefits to the user community through pilot applications or other testing.

Open Source Software

During the Environmental Data Summit, we encountered speakers who, recognizing the realities of governmental investments, advocated for several measures that the State could adopt to follow the evolutionary, rather than the revolutionary, path. Paul Ramsay, an open-source software expert and innovator, offered several compelling arguments in favor of open-source software adoption whenever possible:

1. **The new normal.** "Open-source software and the open-source community," he posited, "is the new normal. It is not going away." The solutions are proven to be effective. The most exciting innovations happen in open-source. The riskier choice in technology management is *not* integrating open-source solutions.
2. **Attracting and retaining talent.** Failure of an enterprise is imminent when talented people depart for better opportunities, particularly when the organization is not able to attract talented replacements. The most talented developers, Ramsay contended, are conversant in open-source technologies. In fact, they embrace open-source solutions precisely because they can implement custom solutions to serve specific needs, rather than implementing turn-key solutions that might fall short of the specified needs.
3. **A hybrid approach is key.** Open-source software is typically able to integrate into proprietary solutions. A period of experimentation with the solution is advisable. The software can be run in parallel to minimize risk of failure and afford functional

[4] http://www.christenseninstitute.org/key-concepts/disruptive-innovation-2/

comparisons. Furthermore, a heterogeneous software infrastructure can also mitigate against the risk of catastrophic cyber-attacks by diversifying the targets, whereas an infrastructure with a uniform profile can be more easily exploited by transgressors. In any event, if the open-source solution proves itself successful, then it can be more broadly adopted.

This incremental approach can minimize disruption to current technology expectations and staffing requirements while also fostering a culture of low-risk experimentation to meet the State's goals. New talent will seek to participate in such experiments. The State will encounter new partners who will co-create solutions with engaged public servants.

Of course, open-source software is typically free or low-cost, but this matter must be weighed against the cost of developing or implementing custom solutions. Licensing costs, in any case, are either minimal or absent, whereas proprietary software's licensing fees can consume a significant portion of a data steward's budget.

Data Federation Evolution

In several key respects, the data federation solution discussed elsewhere in this document can also adhere to an incremental, evolutionary approach. While the financial investment in such a solution will be substantial, data federation shall, for the most part, leave the original data repositories intact to continue to meet their individualized missions, serving their often unique communities and fulfilling their respective requirements. This is an important structural aspect to data federation: the different data repositories (nodes) are not subsumed to a central core. Rather, they contribute to a central virtual repository that bridges across heterogeneous data, making the collective whole appear seamless, but the contributing data repositories continue their work unchanged. Further, as data federation progresses, existing and emerging data integration and visualization tools (open source and proprietary) can be used to conduct the science synthesis that is the core objective of this entire effort.

The State would continue to invest in the data collection and processing that produces good data currently found in existing repositories. In fact, the data processing must continue to ensure that aggregation can be performed reliably and precisely. In other words, data federation is additive. It will only replace current data systems to the extent that the State electively seeks to streamline funding.

Revenue/Funding Strategies

Success for the measures that have been proposed in this document is contingent on identifying and acquiring the appropriate form(s) of funding. The development of the business case, mentioned above, will help to characterize many critical ingredients which will, in turn, inform the ideal sources for funding. Possible funding models would include:

Legislated Funding:
Legislate mandatory contribution by participant agencies or the development of a new organization with adequate funding. A budget change proposal could ensure sustained funding and promote the greatest degree of transparency for such a measure.

Sponsorship/Grant Funding:
Funding is granted by foundations, State Bond measures. Money is usually for single projects or a short term without commitment to long-term funding.

Public/Private Partnerships:
A public–private partnership (PPP) is a government service or private business venture which is funded and operated through a partnership of government and one or more private-sector companies. A PPP involves a contract between a public sector authority and a private party, in which the private party provides a public service or project and assumes substantial financial, technical and operational risk in the project. In projects that are aimed at creating public goods, as in the infrastructure sector, the government may provide a capital subsidy in the form of a one-time grant, to make it more attractive to the private investors.[5]

There are other potential funding models. See Appendix A for a broader list.

Whatever funding model is adopted, the chances of success increase with the infusion of new funding sources. Working within the existing budgetary footprint could yield some precursors to success related to preparation and preliminary discussions, but much of the work before us will require additional funding.

The Work Ahead

To effect a cultural shift in California that would incentivize data sharing, California should develop and implement data management plans for all data acquired that clearly incentivize data-sharing. California should tie future funding opportunities to data transparency, similar to the National Institutes of Health and the National Science Foundation's present policies, such that the requisite time to post data are clearly defined. Recognizing that some data must necessarily be restricted at least for specified time (for example due to litigation or implications for sensitive or endangered resources), data-sharing policies should be clearly articulated with reference to state and federal laws as appropriate. California must strategically position its data management plans toward national and international initiatives and standards. Consulting contracts related to data generation should also be subject to these guidelines.

[5] http://en.wikipedia.org/wiki/Public%E2%80%93private_partnership

Remarkably, in Europe, the INSPIRE program is well underway in its mandate to create a comprehensive spatial data federation.[6] Multiple agencies across the EU established an MOU that facilitated the development of this project with detailed timelines, regulations, and promised features, all designed to promote scientific and public understanding of policies and activities that exert an influence on the environment. If the European Union, with its diverse languages, cultures, and nationalities, can achieve such an unified outcome, why cannot California?

The development of a sustainable business model for California's environmental data management is admittedly challenging, but the primary obstacles are not chiefly technological but a crisis of business process and collective will. The data inventory audits, stakeholder analysis and cost benefit analysis are just a few of the initial steps needed to advance a new business model. Key to success will be ensuring that the State's internal stakeholders and leadership recognize the value in this substantial undertaking. They must serve as the thought leaders who champion these strides.

Conclusion

This document articulates a vision for enhancing California's environmental data management to keep pace with the rate of technological and environmental change. We have examined the challenges primarily through the lenses of data stewardship, data visualization, and sustainable business models, all of which require a new and enhanced level of collaboration among scientists, agencies and data providers.

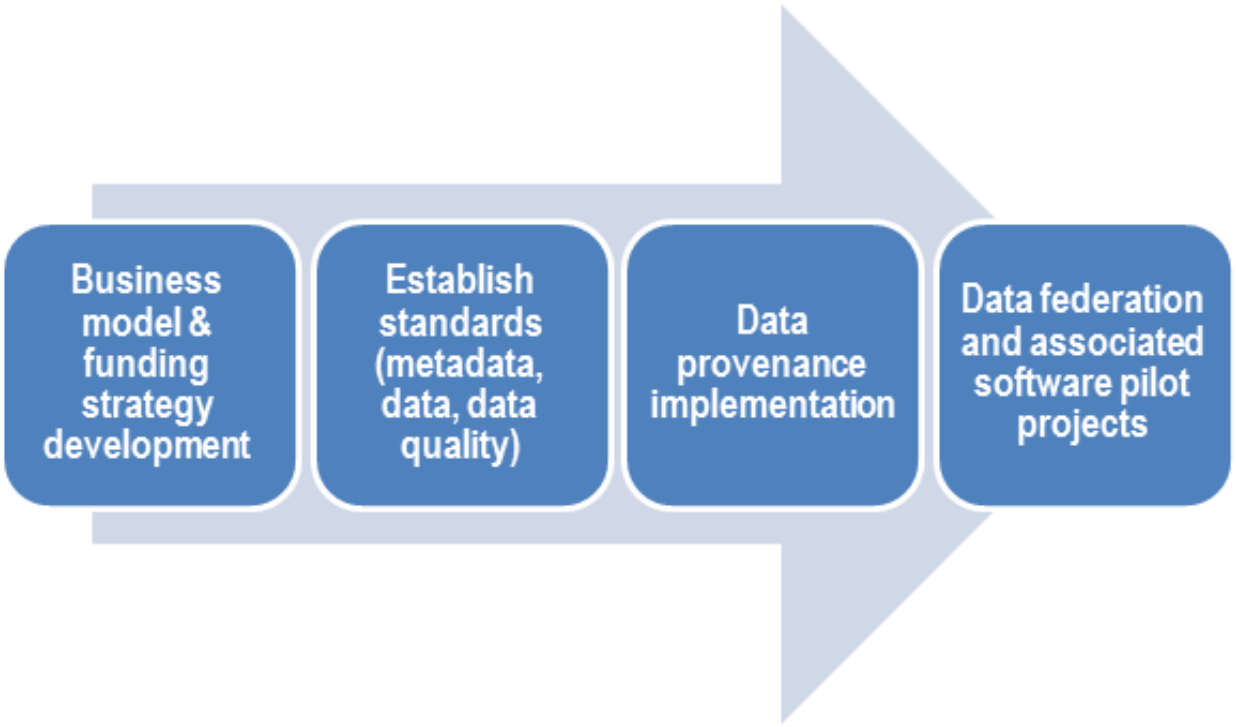
In the course of this collective effort, we have promoted the benefits of data federation, data and development standards, data provenance, and a process for developing sustainable business models, noting their respective capacity to expand the possibilities for data providers and data consumers alike. The challenges that lie in the path of implementation are formidable, and we have taken some effort to catalog those obstacles in each section, but there will still be as-yet-unknown roadblocks to a sustainable solution. While we conduct a pursuit of consensus and broad support for these measures, we realize that we do not have the luxury of time, for waiting runs the risk of falling farther behind. In the meantime, therefore, how can we possibly demonstrate the value of our solutions?

One potential bridging technology resides at a national level. The National Science Foundation supports a project called XSEDE (Extreme Science and Engineering Discovery Environment). A $121-million enterprise, the project partners with scientists and engineers to process data, develop new tools, and accelerate analysis within a high-performance computing environment. The environment offers robust visualization through a portal-based visualization service. This solution would not supplant but rather complement California's existing resources. Nor would it offer a "single bullet" solution to California's challenges, but it would help in developing proof-of-

[6] http://inspire.ec.europa.eu/

concept pilots. Conceivably, a critical mass of California's public agencies or partner organizations might apply for resource enhancement through a creative partnership with the XSEDE project. Similarly, more locally, California's Strategic Growth Council is developing Data Basin as a data aggregation model, and the Delta Restoration Network is pursuing data integration through its own landscape vision framework. These and many other opportunities await.

Along a parallel path, we must pursue an aggressive timeline for comprehensive solution implementation. Here, we have taken an opportunity to synthesize and sequence the recommendations:



In an earlier section, we described the work ahead to develop a sustainable business model. The funding strategy and business case, once shared with strategic partners, will inspire the collaboration and cooperation necessary to motivate the effort. Meanwhile, the work to document current standards and advocate broader adoption continues under the stewardship of the California Water Quality Monitoring Council. Data provenance measures are closely related to the implementation of standards, but treating it as a separate series of tasks, worthy of its own phase, will garner the necessary attention from stakeholders. Now, by linking standards to federation, conferring a broader context and goal for this foundational work, this vision will accelerate the implementation of standards. In other words, the ends will *incentivize* the means.

Which organization, which agency should serve as the standard bearer for this movement? We recommend a partnership-based approach to ensure maximum participation and willing cooperation. This task force or task forces, per se, must be collectively knowledgeable in the challenges articulated in this document and adequately resourced to carry out the work specified. Key decisions -- whether a single or multiple task force can undertake the recommendations, existing partnerships can be leveraged, or new partnerships formed -- must be determined as the first step in this process.

In any event, more robust data sharing among fellow data consumers resulting in accelerated knowledge discovery remains the ultimate goal. Federation, along with the associated steps outlined in this paper, offers the most efficient vehicle to advance this goal. It can be implemented following a pathway that leads to smart sustainability, effective collaboration, and clear standards. Cooperation, collaboration, non-governmental partnerships, and interagency relationships will form the core of our success. Accordingly, our advocated solutions build upon the foundation of the State's established infrastructure. And our success will be measured by our collective advancement.

Appendix A

The following list represents descriptions of various revenue/funding models possibly applicable to State technology initiatives:

Product Based:
In this strategy, each product project provides the funding that is needed for the datasets or use of datasets required in their product. While product-specific funding is the predominant means of funding product development, it may be used to fund core fundamental data development as well. Although product projects are often viewed (and properly so) as a primary source of funding, they may be reluctant to pay for aspects of the core fundamental data operations that they feel are the responsibility of someone else.

Portfolio of Projects:
Multiple groups agree to form and alliance (MOU) and jointly fund the cost of developing a resource to be used by all, pooling data investments across a portfolio of projects. Demonstrate reuse of the data to illustrate “value” -- e.g. four restoration projects, two research projects need x data. Projects would share costs of making data available.

Fee/Usage Based Funding:
Charge a fee proportional to their usage of the core assets. This strategy is similar to enacting a license fee for using a commercial product. Charging such fees is one possible way of obtaining funds for sustaining a program/product.

Taxing of participating projects:

This strategy involves funding selected elements of the product line by levying a tax on each participant/stakeholder. This taxing strategy can use a flat tax or a prorated tax that is based on some particular product attribute (such as product funds, project size, or estimated number of lines of code). The "product-side tax on customers" and "fee based on core asset usage" strategies described here can be viewed as special cases of a taxing strategy.

Legislated Funding:
Legislate mandatory contribution by participant agencies or the development of a new organization with adequate funding. A budget change proposal could ensure sustained funding and promote the greatest degree of transparency for such a measure.

Technology Innovation Fund:
State sponsored investment in projects with uncertain costs and benefits. Examples of this approach are experimental, with adequate incubation and a problem-solving orientation. They are typically shielded from a multi-stakeholder process except during application phase.

Examples of such an approach include:

- The Victorian Government Fund
- Michigan Seed Fund
- Texas Emerging Technology Fund

Sponsorship/Grant Funding:
Funding is granted by foundations or State Bond measures. Money is usually for single projects or a short term without commitment to long-term funding.

Corporate Funding:

This strategy is based on having a corporate-level/program sponsor fund elements of the project -- eg, server infrastructure. In kind donation for corporate value can accrue added benefits.

Public/Private Partnerships:
A public–private partnership (PPP) is a government service or private business venture which is funded and operated through a partnership of government and one or more private-sector companies. A PPP involves a contract between a public sector authority and a private party, in which the private party provides a public service or project and assumes substantial financial, technical and operational risk in the project. In some types of PPP, the cost of using the service is borne exclusively by the users of the service and not by the taxpayer. In other types (notably the private finance initiative), capital investment is made by the private sector on the basis of a contract with government to provide agreed services and the cost of providing the service is borne wholly or in part by the government. Government contributions to a PPP may also be in-kind (notably the transfer of existing assets). In projects that are aimed at creating public goods, as in the infrastructure sector, the government may provide a capital subsidy in the form of a one-time grant, to make it more attractive to the private investors. In some other cases, the government may support the project by providing revenue subsidies, including tax breaks or by removing guaranteed annual revenues for a fixed time period.[7]

Prorated Cost Recovery:

The object of this strategy is to have the projects that have benefited from the product line pay back their fair share of the costs of any software development efforts or services that the product line organization performed on their behalf. This strategy could be extended to include prorating all of, or just elements of, the total cost of sustaining product line operations among the participating project/product developers.

Infrastructure Provision:
Re-Classify data and knowledge as infrastructure for the State of California. Request money from different government funding sources or budgets.

[7] http://en.wikipedia.org/wiki/Public%E2%80%93private_partnership

References

Introduction

United States Geological Survey (Advisory Committee on Water Information). 2014. Charter. http://acwi.gov/a2014_charter.pdf

National Research Council of the Academies. 2012. Sustainable Water and Environmental Management in the California Bay-Delta. Washington, DC: The National Academies Press. http://www.nap.edu/openbook.php?record_id=13394&page=R1

California Natural Resources Agency. 2014. California Water Action Plan. http://resources.ca.gov/california_water_action_plan

Data Stewardship

Delta Stewardship Council. 2014. Environmental Data Summit. http://environmentaldatasummit2014.deltacouncil.ca.gov

Cyberinfrastructure Vision for 21st Century Discovery, *National Science Foundation Cyberinfrastructure Council, March 2007.*

Data Stewardship: An Actionable Guide to Effective Data Management and Data, *David Plotkin, 2014.*

Eckerson, Wayne. “Keys to Creating an Enterprise Data Strategy.” http://www.b-eye-network.com/blogs/eckerson/archives/2011/05/keys_to_creatin.php

Ghosh, S. 2011. Transparent and Commercialized?: Managing the Public-Private Model for Data Production and Use. Managing the Public-Private Model for Data Production and Use (March 7, 2011). Univ. of Wisconsin Legal Studies Research Paper, (1155). http://works.bepress.com/context/shubha_ghosh/article/1005/type/native/viewcontent

Institute of Network Cultures. “Beyond Distributed and Decentralized: What is a Federated Network?” http://networkcultures.org/unlikeus/resources/articles/what-is-a-federated-network/

Master Data Management and Customer Data Integration for a Global Enterprise, Chapter 6: Data governance and data stewardship strategies and best practices, *Alex Berson and Larry Dubov,2007.*

Open Geospatial Consortium. http://www.opengeospatial.org/ogc/markets-technologies/environment-natural-resources

Oracle Corporation. “Best Practices for Real-Time Data Warehousing,” *An Oracle White Paper, 2014*.

Rausch, Nancy and Stearn, Tim. “Best Practices in Data Integration: Advanced Data Management.” SAS Institute, Cary, NC, 2011.

Socrata. “Open Data Implementation in Six Steps.” http://www.socrata.com/open-data-field-guide-chapter/implementation/#nine

Data Visualization

Gamble, Paul and Blackwell, John. 2002. Knowledge Management: A State of the Art Guide. P. 43.

Haden, V.R., M.T. Niles, M. Lubell, J. Perlman, and L.E. Jackson. 2012. Global and local concerns: What attitudes and beliefs motivate farmers to mitigate and adapt to climate change? PLoS ONE 7(12): e52882. doi:10.1371/journal.pone.0052882

Kosara, Robert, and Jock Mackinlay, (2013), “Storytelling: The Next Step for Visualization” Computer (Special Issue on Cutting-Edge Research in Visualization), 46(5): 44 – 50

Lubell, M., V. Hillis, and M. Hoffman. 2011. Innovation, cooperation, and the perceived benefits and costs of sustainable agriculture practices. Ecology and Society 16(4): 23. http://dx.doi.org/10.5751/ES-04389-160423

Robinson AH, Safran SM, Beagle J, Grossinger RM, Grenier JL, Askevold RA. 2014. A Delta Transformed: Ecological Functions, Spatial Metrics, and Landscape Change in the Sacramento-San Joaquin Delta. Richmond, CA: San Francisco Estuary Institute - Aquatic Science Center. http://www.sfei.org/documents/delta-transformed-ecological-functions-spatial-metrics-and-landscape-change-sacramento-san

Shilling, F.M. 2012. Lower Sacramento River 2011 water quality report card. Prepared for the Sacramento River Watershed Program. Pp. 40.

Shilling, F.M. 2013. The California water sustainability indicators framework, Phase II: State and regional pilots. Report to Department of Water Resources. Pp. 311.

Tufte, Edward, (2001 [1983]), The Visual Display of Quantitative Information, 2nd ed. (First edition 1983). Cheshire, CT: Graphics Press.

Sustainable Business Models

Carlin, Allan. U.S. EPA. “The New Challenge to Cost-Benefit Analysis.” http://object.cato.org/sites/cato.org/files/serials/files/regulation/2005/9/v28n3-3.pdf

Cost-Benefit Analysis and the Environment. http://www.oecd.org/greengrowth/tools-evaluation/36190261.pdf

Delen, D. and Demirkan, H. (2013), “Data, information and analytics as services”,Decision Support Systems, Vol. 55, No. 1, pp. 359–363.

Doms, Mark. “The Commerce Department’s Strategic Plan: The Value of Government Data.” http://www.commerce.gov/blog/2014/03/24/commerce-department%E2%80%99s-strategic-plan-value-government-data

Eppler, Martin J and Helfert, Markus. A Classification and Analysis of Data Quality Costs. http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202004/Papers/AClassificationandAnalysisofDQCosts.pdf

Jaffe, A., Newell, R. and Stavins, R. (2004), “A Tale of Two Market Failures: Technology and Environmental Policy

Orth, K., R. Robinson, and W. Hansen. 1998. “Making More Informed Decisions in Your Watershed When Dollars Aren't Enough.” IWR Report 98-R-1. U.S. Army Corps of Engineers, Alexandria, Virginia.

Zott, C., Amit, R. and Massa, L. (2011), “ The Business Model : Recent De velopments and Future Research”, Journal of Management , Vol. 37, No. 4 , pp. 1019 – 1042.